



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Optical Flow-Guided Generative Adversarial Network For Video Super-Resolution

Ketaki Sawashe^{1*}, Bageshree Pathak², Supriya Mangale³

^{1,2,3} MKSSS's Cummins College of Engineering for Women, affiliated to Savitribai Phule Pune University, Pune, India.

*Correspondence: ketaki.sawashe@cumminscollege.in

Abstract

Video Super-Resolution has recently gained significant attention due to its importance in surveillance, medical imaging, multimedia streaming and video restoration applications. However, reconstructing high-resolution video frames from low-resolution sequences remains challenging because the network must recover fine spatial details while also preserving temporal consistency between consecutive frames. This paper presents a Flow-Guided Generative Adversarial Framework for Video Super-Resolution that integrates optical flow-based temporal alignment, multi-frame feature fusion, and adversarial perceptual learning to enhance video reconstruction quality. The proposed framework employs a pretrained RAFT network to estimate inter-frame motion and align neighboring frames to a reference frame. The aligned frames are subsequently fused and processed through a residual convolutional reconstruction network to generate high-resolution outputs. To improve perceptual quality and texture fidelity, a PatchGAN discriminator is incorporated together with perceptual and adversarial loss functions. The framework achieves a PSNR of 28.58 dB and SSIM of 0.8672 on REDS validation data, while obtaining 25.42 dB PSNR and 0.8253 SSIM on VID4. Ablation studies further confirm the contributions of flow-guided alignment, multi-frame feature fusion, and adversarial perceptual learning to overall reconstruction performance. Under a fixed, modest training budget, the proposed framework achieves competitive perceptual quality and structural similarity relative to larger state-of-the-art architectures, indicating a favorable trade-off between reconstruction accuracy and perceptual realism for lightweight video super-resolution.

Keywords: Video Super-Resolution, Generative Adversarial Networks, Optical Flow, Flow-guided Alignment, Deep Learning.

1. Introduction

Video Super-Resolution (VSR) aims to reconstruct a High-Resolution (HR) frame from its Low-Resolution (LR) counterpart by using both spatial information within individual frames and temporal correlations across adjacent frames [1]. This is different from the Single-Image Super Resolution (SISR), where recovering the high frequency details is the major task rather than temporal alignment. VSR considers sequence of frames that are inherently coupled through motion, scene dynamics and temporal consistency. The positions of objects change from frame to frame. This nature of VSR makes it more challenging than SISR. The errors introduced in one frame can propagate as temporal artifacts in upcoming frames, leading to visually disturbing occurrences like flickering, ghosting or motion inconsistency. Hence, a successful VSR model should address both the objectives of spatial detail recovery along with temporal consistency.

One of the main challenges in VSR comes from complex and non-uniform motion in the real-world videos. Objects may undergo large displacements, non-rigid deformations, occlusions and disocclusions, which further complicates the task of leveraging temporal information effectively. Using temporal redundancy provides additional information, which makes it possible to create higher-quality images than looking at a single frame alone. However, this only works with accurate alignment of frames. Even small misalignments can lead to severe degradation in high-frequency regions. Traditional VSR methods depended on explicit motion estimation and compensation using optical flow, to align neighboring frames before fusion [2-4]. However, inaccuracies in motion estimation, specifically near occlusion boundaries or fast motion resulted in visual artifacts and accumulation of errors. These limitations have motivated the development of learning-based approaches that can efficiently model the temporal dependencies.

In recent years, Deep Learning (DL) based approaches have significantly advanced the state of VSR by enabling end-to-end learning of spatio-temporal representations. Early Convolutional Neural Network (CNN) based methods improved the quantitative performance over the traditional techniques by using frame alignment modules and temporal aggregation strategies [5]. However, CNN-based methods lack the ability to model long range spatial and temporal dependencies. As a result, these models have trouble in understanding fast motions

and long-term temporal relationships, which are common in videos. Transformer-based architectures have emerged as an efficient method for VSR because of their ability to model long range dependencies using self-attention mechanisms. Models such as the Video Restoration Transformer (VRT) [6] have demonstrated remarkable performance improvements in the metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). These models look at spatial and temporal dimensions simultaneously and aggregates information from distant frames to handle fast motions and complex scenes. The hierarchical and window-based attention modules help in focusing on small, local areas while also looking at the global context, thus recreating high-quality images efficiently. Despite efficiency in quantitative measures, Transformer-based VSR models show limitations in terms of perceptual quality. Their optimization objectives are typically dominated by pixel-wise or distortion-based loss functions such as L1 or L2 loss. These losses are effective for improving PSNR and SSIM by minimizing the average error, but they fail to recover perceptually realistic textures. As a result, Transformer-based VSR outputs often lack sharp edges, fine textures and realistic high-frequency details. This limitation is of major concern in regions with repetitive patterns, fine structures or complex textures, where perceptual quality plays a crucial role in human visual assessment. In short, standard distortion metrics don't always align with human perception, a conflict known as the perception-distortion tradeoff. Models optimized for low distortion (high PSNR/SSIM) tend to produce perceptually smooth results, whereas models optimized for perceptual realism often sacrifice distortion metrics. This tradeoff is further complicated by temporal constraints in VSR. Perceptual enhancement of individual frames can improve the visual quality, but presence of inconsistencies may lead to distracting artifacts when played as a video. Therefore, the VSR models need to find the right balance between spatial realism and temporal stability.

Generative Adversarial Networks (GANs) are widely used to improve image super-resolution by addressing perceptual limitations and producing realistic textures through adversarial training [7]. The model trains generator to fool the discriminator, while the discriminator's task is to recognize the real and generated HR images. GAN based models can produce sharper and more visually appealing results than that of purely distortion optimized models [8]. These models have shown significant improvement in the perceptual quality in case of image SR. however, applying GANs to VSR introduce several challenges. Adversarial training, which is inherently unstable, gets amplified when applied to videos because of temporal connection of frames. This may lead to frame-to-frame variation causing flickering artifacts. Even when spatio-temporal discriminators are used, the adversarial training can amplify minor inconsistencies across frames, making stable training difficult. Also, GANs may produce visually plausible but temporally inconsistent textures which are not present in the original frame. As a result, GAN based VSR models achieve better perceptual quality but poor temporal stability, which limits their practical application.

Existing VSR methods often face a tradeoff between temporal consistency and perceptual quality. While flow-guided methods provide accurate alignment, GAN-based methods improve visual realism. Combining these strengths remains an important research direction. In this context, flow-guided modeling acts as a crucial link between multi frame feature fusion and GAN-based perceptual enhancement. It uses optical flow to explicitly map the inter-frame motion, thus ensuring proper feature alignment over time. Integrating flow guidance into the GAN framework enables adversarial learning to utilize motion-aligned representations, which reduces temporal ambiguity and minimizes flickering artifacts. By integrating flow-guided adversarial supervision, the model can enhance perceptual details and also maintain temporal consistency.

Comparing VSR methods is difficult because many things change at the same time. Different papers use different network sizes, different training schedules, different amounts of training data, different alignment methods and different loss functions. So when one method reports a higher PSNR, it is not clear which choice caused the gain. This work does not claim the best PSNR or SSIM. Instead, it studies how each part of the framework affects the result, while keeping the network size, training schedule and evaluation settings the same in every experiment.

The main contributions are:

- A flow-guided GAN framework that uses a frozen pretrained RAFT model, so that alignment errors do not mix with adversarial training.
- A component-wise ablation that shows how much temporal alignment, multi-frame fusion and adversarial learning each add, under the same training settings.
- A staged training schedule in which LPIPS and GAN losses are added only after a distortion-based warm-up, and adversarial loss is applied only to the center frame to keep training stable.
- A study of the trade-off between pixel accuracy and visual quality in a small model, reporting both distortion and perceptual metrics instead of PSNR alone.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed framework. Section 4 describes the experimental setup and results. Section 5 discusses findings and future work. Finally, Section 6 concludes the paper.

2. Related Work

Early DL approaches for VSR used CNNs to exploit temporal redundancy across consecutive frames. One of the earliest and most influential works, VESPCN by Caballero et al. [2], introduced a real-time framework that jointly performs optical flow estimation, motion compensation and spatial super-resolution within a unified CNN architecture. By explicitly aligning adjacent frames using estimated motion, VESPCN demonstrated that temporal information could significantly improve reconstruction quality over single-image SR. Similar flow-based alignment strategies were further explored in Sub-Pixel Motion Compensation (SPMC) [3], proposed to better handle fractional displacements and in Frame-Recurrent Video Super-Resolution (FRVSR) [4], which adopted a frame-recurrent formulation to propagate HR estimates over time. Beyond explicit flow, several works investigated implicit motion modeling through dynamic filtering or recurrent structures. Jo et al. [5] proposed dynamic upsampling filters that adapt spatial kernels based on local motion patterns without explicit motion compensation, while recurrent CNN-based methods such as bidirectional recurrent convolutional networks [9], deep recurrent ResNet [10] and frame-recurrent back-projection networks [11] emphasized temporal information propagation through hidden states. These recurrent formulations improved temporal stability and reduced flickering compared to purely feed-forward designs, but remained limited in capturing complex motion over long temporal spans. More recently, BasicVSR [12] introduced a simple yet highly effective propagation framework that demonstrated many sophisticated modules were unnecessary when bidirectional propagation and feature alignment were properly designed. Building upon this framework, BasicVSR++ [13] introduced second-order grid propagation together with improved optical flow alignment, achieving substantial gains on challenging video restoration benchmarks.

A major advance in CNN-based VSR came with the introduction of deformable convolution for alignment. EDVR by Wang et al. [14] proposed enhanced deformable convolutional networks that learn spatially adaptive sampling offsets for multi-frame alignment, achieving state-of-the-art performance across several video restoration tasks. Subsequent works such as TDAN [15], deformable non-local networks [16], deformable 3D convolutions [17] and FDAN [18] further refined deformable alignment mechanisms to better handle complex motion and occlusions. These approaches significantly improved temporal stabilization and distortion metrics such as PSNR and SSIM. Despite their success, CNN-based VSR methods fundamentally rely on local receptive fields and short-range temporal aggregation, even when enhanced with deformable alignment or recurrent propagation. As a result, their ability to model long-range temporal dependencies remains limited, particularly in videos with complex motion patterns or long-term occlusions. Moreover, most CNN-based approaches are optimized using pixel-wise losses (e.g., L1 or Charbonnier), which encourages conservative averaging across frames and leads to over-smoothed reconstructions lacking perceptual detail, a limitation that becomes more apparent at higher upscaling factors.

Motivated by the limitations of local convolutional operators, recent research has explored Transformer architectures for video super-resolution, leveraging self-attention to model long-range spatial and temporal dependencies. Early efforts such as the spatio-temporal transformer network by Kim et al. [19] demonstrated the potential of attention mechanisms for video restoration tasks. Building on advances in vision transformers, subsequent works introduced more specialized architectures for VSR. SwinIR [20], although initially proposed for image restoration, demonstrated the effectiveness of shifted-window attention for capturing multi-scale context with manageable computational cost and has since been adapted as a backbone for video restoration pipelines. Dedicated Transformer-based VSR models further extended these ideas. The Video Super-Resolution Transformer [21], trajectory-aware transformers [22], Real-Time Spatial Temporal Transformer (RSTT) [23], Local-Global Fusion Network (LGFN) [24] and collaborative spatial-temporal transformers such as CTVSR [25] explicitly modeled temporal correlations using attention across multiple frames. These methods showed strong improvements in distortion-oriented metrics, particularly on benchmarks with large motion or long temporal dependencies. The culmination of this line of research is VRT by Liang et al. [6], which proposed a unified Transformer framework capable of handling multiple video restoration tasks through hierarchical spatial-temporal attention and flexible token aggregation. VRT consistently achieves state-of-the-art PSNR and SSIM across standard datasets, demonstrating the strength of attention-based temporal modeling. However, Transformer-based VSR methods are typically trained using distortion-driven objectives, such as L1 or Charbonnier loss, sometimes combined with perceptual losses. While this optimization strategy yields high quantitative scores, it inherently biases the model toward minimizing pixel-wise error across frames, resulting in perceptually smooth and conservative reconstructions. Fine textures, high-frequency details and visually pleasing sharpness are often suppressed, especially in challenging regions with motion ambiguity. This behavior directly reflects the well-known perception-distortion trade-off, where optimizing for distortion metrics does not necessarily align with human visual preferences. Consequently, despite their superior temporal modeling capability, Transformer-based VSR models alone do not fully address perceptual realism.

GANs have been widely adopted in super-resolution to overcome the perceptual limitations of distortion-oriented training. SRGAN by Ledig et al. [26] was the first to demonstrate that adversarial learning combined with perceptual loss can produce visually convincing high-frequency details absent in PSNR-optimized models. ESRGAN [27] further improved perceptual quality through residual-in-residual dense blocks and a relativistic discriminator, setting a new standard for perceptual image SR. More recently, Real-ESRGAN [28] addressed real-world degradations, showing robust perceptual enhancement under practical conditions. Extending GAN-based learning to video, however, introduces significant challenges. Early attempts such as TecoGAN [29] incorporated temporal coherence constraints into adversarial training to reduce flickering artifacts. Subsequent works explored recurrent GAN formulations, including iSeeBetter [30], progressive fusion GANs [31] and recurrent back-projection GANs [32]. While these methods improved perceptual sharpness, they often suffered from temporal instability, manifested as flickering textures or inconsistent details across frames. The root cause lies in the adversarial objective itself: most GAN-based VSR approaches employ frame-wise discriminators or weak temporal constraints, which fail to enforce long-term consistency. Moreover, jointly learning motion estimation and adversarial synthesis is notoriously unstable, as errors in alignment are amplified by the discriminator. As a result, naively applying GANs to video super-resolution often leads to visually appealing single frames but temporally incoherent videos, highlighting the need for more principled integration strategies.

Accurate motion estimation plays a central role in addressing temporal consistency, particularly when perceptual losses are introduced. Classical flow-based VSR methods [2,3] already demonstrated the importance of alignment, but recent advances in deep optical flow estimation have significantly improved motion accuracy. FlowNet [33] and FlowNet2 [34] introduced end-to-end learning for optical flow, while RAFT by Teed and Deng [35] achieved state-of-the-art accuracy through recurrent all-pairs field transforms. Several VSR works have leveraged high-resolution optical flow to enhance reconstruction quality. MEMC-Net [36] unified motion estimation and compensation for video enhancement tasks. More recent approaches emphasize flow-guided alignment as a modular component, decoupling motion estimation from reconstruction to improve training stability. This decoupling is particularly important for GAN-based VSR, where inaccurate alignment can destabilize adversarial learning and exacerbate temporal artifacts. Using the pretrained, high-quality optical flow estimators such as RAFT allows motion modeling to be treated as a reliable prior rather than a learned auxiliary task. This strategy provides a stable temporal backbone for perceptual learning, enabling GAN-based models to focus on texture synthesis while preserving frame-to-frame consistency. Although optical flow has been widely used for alignment, few works integrate high-quality pretrained flow and stable adversarial perceptual learning within a unified framework. The gaps motivate the proposed approach, which combines multi-frame feature fusion with flow-guided alignment and GAN-based perceptual learning, aiming to achieve both high perceptual fidelity and temporal consistency in video super-resolution.

3. Proposed Method

3.1 Overall Framework

Given a sequence of LR frames, the objective of VSR is to reconstruct a temporally consistent HR sequence that preserves both fine spatial details and temporal consistency. We propose a flow-guided GAN-based VSR framework. Rather than directly super-resolving each frame independently, we first align all neighboring LR frames to the reference (center) frame using a pretrained optical flow model. The aligned frames are then fused and passed through a generator network to reconstruct the HR output. Adversarial learning is employed using PatchGAN discriminator to enhance perceptual quality, but is carefully restricted to avoid temporal instability. The overall architecture of this framework is shown in Figure 1.

Let $\{I_{t-k}^{LR}, \dots, I_t^{LR}, \dots, I_{t+k}^{LR}\}$ denote a temporal window of $2k+1$ LR frames centered at time index t . The goal is to estimate the corresponding HR frame I_t^{HR} . The input to this framework is a sequence of five consecutive LR frames:

$$I_{t-2}^{LR}, I_{t-1}^{LR}, I_t^{LR}, I_{t+1}^{LR}, I_{t+2}^{LR}$$

Our method consists of four main stages:

1. Temporal alignment using pretrained optical flow to warp neighboring frames to the reference frame.
2. Multi-frame feature fusion via a CNN-based generator
3. Adversarial learning using a PatchGAN discriminator
4. Perceptual refinement using adversarial and perceptual losses applied in a delayed and controlled manner.

3.2 Flow-Guided Temporal Alignment

Accurate temporal alignment is very important for effective multi-frame fusion. Let I_t^{LR} denote an LR frame and I_t^{LR} be the reference frame. We estimate the optical flow field using a pretrained RAFT-Large optical flow network, which is kept frozen during training.

$$F_{i \rightarrow t} = \text{RAFT}(I_i^{LR}, I_t^{LR}) \quad (1)$$

The flow field maps pixels from frame i to the reference frame t . The aligned frame I_t^{LR} is obtained via backward warping defined at each spatial location x as

$$\tilde{I}_i^{LR}(x) = I_i^{LR}(x + F_{i \rightarrow t}(x)) \quad (2)$$

where $F_{i \rightarrow t}(x)$ denotes displacement vector at position x . For notational simplicity, we represent the warping operation as,

$$\tilde{I}_i^{LR} = W(I_i^{LR}, F_{i \rightarrow t}) \quad (3)$$

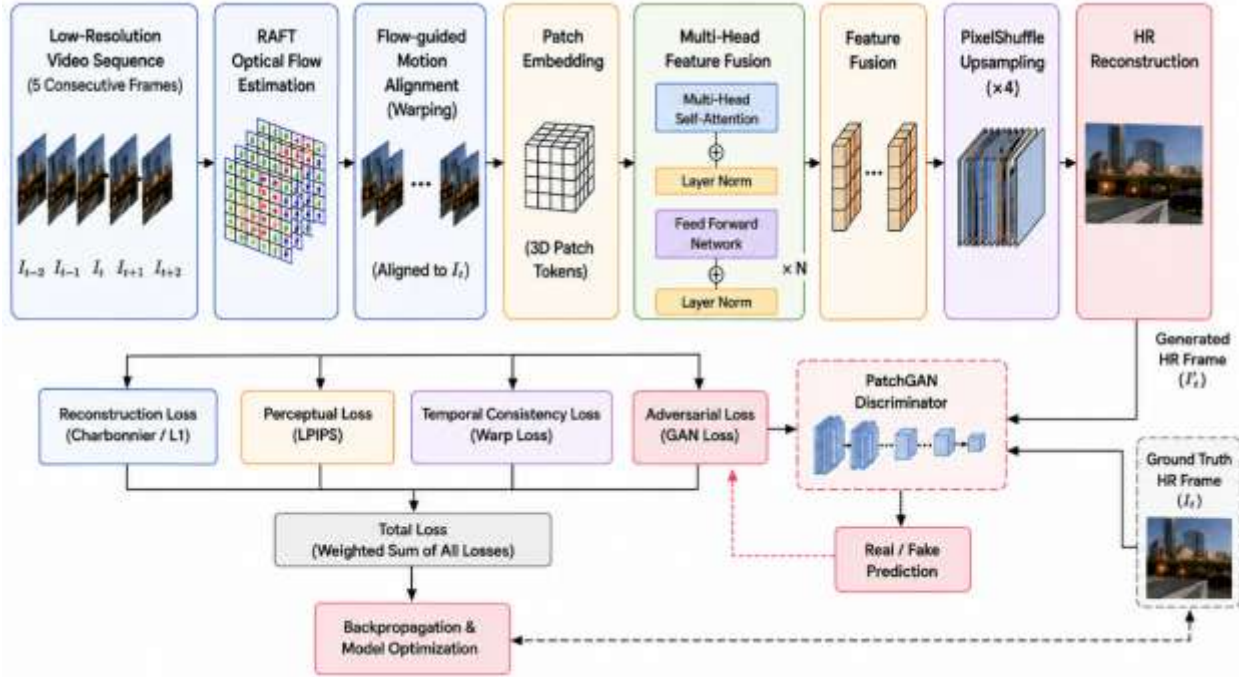


Figure 1. Overall architecture of the proposed framework.

The aligned input sequence is then, $\mathcal{I} = \{\tilde{I}_{t-k}^{LR}, \dots, I_t^{LR}, \dots, \tilde{I}_{t+k}^{LR}\}$ which is subsequently provided to the generator for spatial-temporal fusion. To satisfy RAFT's architectural constraints, input frames are padded to dimensions divisible by 8 prior to flow estimation and unpadded afterward. This step reduces temporal misalignment and improves reconstruction consistency across frames. We intentionally freeze the optical flow network during training. Jointly optimizing motion estimation together with adversarial learning often leads to instability, since inaccuracies in flow can be amplified by GAN gradients. By relying on a strong pretrained RAFT model, motion is treated as a reliable prior rather than an additional variable to learn. This results in more stable optimization and allows perceptual objectives to focus on restoring details instead of correcting alignment errors.

3.3 Generator Architecture

The generator G takes the aligned LR sequence $\mathcal{I}$ and predicts the HR center frame:

$$\hat{I}_t^{HR} = G(\mathcal{I}) \quad (4)$$

Instead of explicit attention layers, temporal aggregation is achieved through explicit alignment followed by channel-wise fusion.

3.3.1 Input Fusion

The aligned LR frames are concatenated along the channel dimensions, $X = \text{Concat}(\tilde{I}_{t-k}^{LR}, \dots, I_t^{LR}, \dots, \tilde{I}_{t+k}^{LR})$, resulting in a 15-channel input tensor.

3.3.2 Feature Extraction and Refinement

A convolutional head projects the concatenated input into a higher-dimensional feature space, followed by eight residual blocks, each consisting of two 3×3 convolutional layers with ReLU activation and residual connections. This design enables effective spatial feature refinement while maintaining training stability.

3.3.3 Up sampling and Reconstruction

Spatial resolution is increased by a factor of 4 using PixelShuffle-based upsampling, followed by a final convolutional layer. A tanh activation is applied to normalize output pixel values to the range $[-1,1]$, consistent with input normalization.

3.3.4 Discriminator Architecture

We adopt a PatchGAN-style discriminator D , which outputs a spatial map of real/fake predictions rather than a single scalar. A PatchGAN-style discriminator is employed to provide adversarial supervision. The discriminator consists of 3a sequence of strided convolutional layers with increasing channel depth, interleaved with LeakyReLU activations and batch normalization. Instead of producing a single realism score, the discriminator outputs a spatial map of authenticity scores, enabling localized texture discrimination. Adversarial learning is applied only to the super-resolved output corresponding to the center frame, which significantly improves training stability and reduces temporal flickering. Applying adversarial loss to all frames in a video sequence often leads to temporal flickering, as the discriminator lacks awareness of motion coherence. By restricting adversarial supervision to the center frame only, we ensure that perceptual sharpness is enforced without propagating inconsistent gradients across time. Temporal consistency is instead preserved through flow-guided alignment and distortion losses.

3.3.5 Loss Functions

The generator is trained using a weighted combination of reconstruction and perceptual losses.

Reconstruction Losses: During early training, the generator is supervised using a combination of Charbonnier loss and L1 loss in order to enforce accurate structure and temporal stability.

$$L_{charb} = E [(\hat{I}_t^{HR} - I_t^{HR})^2 + \epsilon]^{1/2} \quad (5)$$

$$L_{L1} = E [|| \hat{I}_t^{HR} - I_t^{HR} ||] \quad (6)$$

Perceptual Loss: After a warm-up phase, perceptual supervision is introduced using LPIPS loss computed with a VGG-based pretrained network. This loss improves texture realism and visual sharpness.

$$L_{LPIPS} = E [|| \phi(\hat{I}_t^{HR}) - \phi(I_t^{HR}) ||] \quad (7)$$

Adversarial Loss: The adversarial objective is implemented using a binary cross-entropy formulation:

$$L_{GAN} = E [\log D(I_t^{HR})] + E [\log(1 - D(\hat{I}_t^{HR}))] \quad (8)$$

The adversarial loss is applied with a small weight to prevent instability and over-sharpening.

Complete Loss: The complete generator loss is defined as:

$$L_G = L_{charb} + 0.5 L_{L1} + \lambda_1 L_{LPIPS} + \lambda_2 L_{GAN} \quad (9)$$

where λ_1 and λ_2 are activated only after the warm-up stage.

3.3.6 Training Strategy

Phase 1, Distortion-Oriented Warm-Up: During the first T epochs, only distortion losses are applied.

$$L_G = L_{charb} + 0.5 L_{L1} \quad (10)$$

This allows the generator to learn accurate structure and motion-consistent reconstruction before any adversarial pressure.

Phase 2, Perceptual Fine-Tuning: After warm-up, LPIPS and GAN losses are gradually introduced. Introducing GAN loss too early destabilizes training, especially when reconstruction is still inaccurate. Delayed adversarial learning ensures that the discriminator refines details rather than correcting large structural errors.

4. Experiments And Results

4.1 Datasets

This model is evaluated on two benchmark datasets: REDS (REalistic and Dynamic Scenes) and VID4. The REDS dataset is used for training, while its validation part (REDS_val) is used for quantitative evaluation. REDS contains high-quality video sequences with complex motion and realistic degradations, making it suitable for learning and testing video super-resolution models. In addition, the VID4 dataset is used to evaluate the generalization capability of the model. VID4 consists of four standard sequences (Calendar, City, Foliage and Walk), which are widely used in the literature for benchmarking VSR methods due to their diverse motion patterns and textures.

4.2 Implementation

The model is implemented in PyTorch and trained on a GPU-enabled system. A sequence of five consecutive LR frames is used as input to capture temporal information and the network predicts the corresponding high-resolution center frame. The overall training objective consists of multiple loss components:

- Reconstruction loss (Charbonnier + L1) to ensure pixel-level accuracy
- Perceptual loss (LPIPS) to improve visual quality
- Adversarial loss (binary cross-entropy) to enhance realism

- Temporal consistency loss to enforce coherence between frames

A staged training strategy is adopted, where the model is initially trained without adversarial loss to stabilize learning and the GAN component is introduced in later epochs.

The model performance is evaluated using three standard metrics:

- PSNR (Peak Signal-to-Noise Ratio) for reconstruction fidelity.
- SSIM (Structural Similarity Index) for structural similarity.
- LPIPS (Learned Perceptual Image Patch Similarity) for perceptual quality.

4.3 Result Analysis

Table 1 presents the quantitative comparison of the proposed model with several existing VSR approaches. The values for the other methods are taken from their original papers. Those models are much larger, were trained for longer and used the full training set. So, the table shows where the proposed method stands compared to published results, but it does not explain which design choice causes the difference. The controlled comparison in this work is Table 2, where every configuration uses the same generator, the same data split, the same training schedule and the same evaluation settings. The experiments reported here were run under a limited training budget. Large-scale frameworks such as VRT and BasicVSR++ are usually trained for several hundred thousand iterations, often on more than one dataset. The difference in training budget should be kept in mind when reading Table 1, since the reported values reflect both the architecture and the training scale. Although the proposed framework does not outperform extremely deep transformer-based architectures such as VRT in terms of PSNR, the model demonstrates competitive structural similarity and perceptual reconstruction capability using a comparatively lightweight architecture. The comparatively strong SSIM values obtained on both datasets demonstrate the capability of the model to preserve structural consistency across video frames. Moreover, the 0.178 LPIPS score suggest that adversarial and perceptual optimization significantly improve visual realism and texture sharpness.

Figure 2 presents representative qualitative reconstruction results on the Calendar and Walk sequences from the VID4 dataset. The reconstructed outputs are visually closer to the ground-truth. The integration of adversarial learning further enhances perceptual quality by encouraging the generator to synthesize realistic textures rather than overly smooth reconstructions. As a result, the generated outputs exhibit improved visual sharpness and better perceptual consistency across consecutive frames.

Table 1. Quantitative comparison (average PSNR/SSIM) with state-of-the-art methods for video super-resolution ($\times 4$) on REDS_val and VID4.

Method	REDS_val	VID4
Bicubic	26.14 / 0.7292	23.78 / 0.6347
SwinIR [20]	29.05 / 0.8269	25.68 / 0.7491
TOFlow	27.98 / 0.7990	25.89 / 0.7651
RBPN [11]	29.63 / 0.8502	26.73 / 0.8029
EDVR [14]	31.09 / 0.8800	27.35 / 0.8264
BasicVSR [12]	31.42 / 0.8909	27.24 / 0.8251
VSRT [21]	31.19 / 0.8815	27.36 / 0.8258
VRT [6]	32.19 / 0.9006	27.93 / 0.8425
LGFN [24]	-	27.31 / 0.8248
DUF [5]	-	27.34 / 0.8327
Ours	28.58 / 0.8672	25.42 / 0.8253

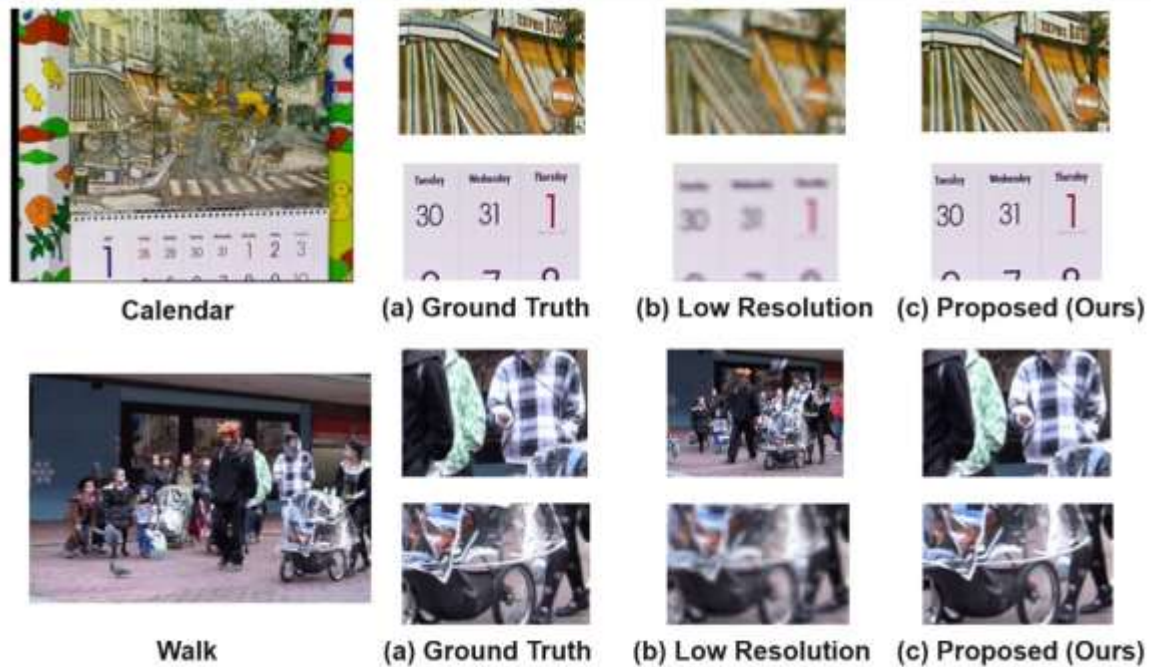


Figure 2. Qualitative super-resolution results of the proposed methos on the Calendar and Walk sequences from the VID4 dataset. Source: Authors, (2026).

4.4 Ablation Study

An ablation study was conducted to analyze the contribution of major components within our framework. Different model configurations were evaluated by selectively removing adversarial learning and temporal fusion modules. The baseline CNN generator produced comparatively smoother outputs with reduced texture sharpness. Incorporating temporal fusion improved structural consistency by effectively utilizing neighbouring frame information. This improved spatial detail reconstruction and structural consistency across video frames. Finally, the addition of adversarial optimization significantly improved perceptual quality and texture realism, leading to lower LPIPS values and visually sharper reconstructions. The qualitative observations align with the obtained LPIPS results, confirming that the proposed framework effectively improves perceptual reconstruction quality while maintaining temporal consistency.

Table 2. Ablation study on the REDS_val dataset.

CNN Feature Extractor	RAFT-based Temporal Alignment	Multi-frame Feature Fusion	GAN-based Perceptual Learning	PSNR ↑	SSIM ↑	LPIPS ↓
✓	×	×	×	26.91	0.8214	0.243
✓	✓	×	×	27.54	0.8421	0.216
✓	✓	✓	×	28.11	0.8578	0.192
✓	✓	✓	✓	28.58	0.8672	0.178

5. Discussion And Future Work

The experimental results demonstrate that the proposed method is capable of generating visually improved HR video frames while maintaining satisfactory temporal consistency. Unlike conventional CNN-based VSR methods that primarily focus on local feature extraction, the proposed approach benefits from multi frame fusion together with flow-guided temporal alignment and adversarial perceptual optimization. The integration of these components enables the network to reconstruct sharper textures and preserve structural details more effectively across consecutive frames. The quantitative results on REDS_val and VID4 indicate that the proposed method achieves competitive SSIM and LPIPS performance, particularly in terms of perceptual reconstruction quality. Although the PSNR values remain lower than very deep state-of-the-art models such as VRT, the proposed framework demonstrates a balanced tradeoff between distortion-oriented reconstruction and perceptual realism. This behavior is expected because adversarial learning generally prioritizes visual

sharpness and texture synthesis rather than maximizing pixel-wise similarity. Consequently, the generated frames appear visually more detailed and less over-smoothed compared to purely distortion-optimized methods. The qualitative observations further support this conclusion. The reconstructed outputs preserve edge sharpness, fine textures and structural continuity across frames. The use of pretrained RAFT-based alignment significantly reduces motion inconsistency and improves temporal coherence, especially in regions containing moderate motion. Moreover, the staged training strategy contributes to stable adversarial optimization by allowing the generator to first learn accurate structural reconstruction before introducing perceptual supervision.

The ablation study confirms the importance of each component within the framework. RAFT-based temporal alignment improves frame consistency; the multi-frame feature fusion module enhances contextual feature representation and GAN-based perceptual learning contributes substantially to visual realism. The reduction in LPIPS after introducing adversarial optimization indicates that the generated outputs become perceptually closer to ground-truth HR frames. Despite these improvements, several limitations remain. First, the current framework employs a relatively lightweight generator architecture in order to maintain training stability and manageable computational complexity. As a result, the reconstruction capability remains lower than extremely deep hierarchical architectures. Second, the present method performs frame-level adversarial supervision only on the center frame, which improves stability but limits explicit long-term temporal adversarial learning. Third, although pretrained optical flow improves alignment quality, inaccuracies in motion estimation may still affect reconstruction performance in scenes containing large displacement, occlusion, or fast non-rigid motion.

Future work will focus on addressing these limitations through several possible directions. One potential extension is the integration of hierarchical window-based transformers such as Swin Transformer or VRT-inspired modules to improve long-range temporal modeling. Another promising direction involves incorporating deformable attention mechanisms for adaptive temporal alignment instead of relying solely on explicit optical flow warping. Additionally, introducing spatio-temporal discriminators capable of jointly evaluating perceptual realism and temporal consistency may further reduce flickering artifacts and improve video quality.

6. Conclusion

In this paper, a flow-guided GAN framework has been presented for VSR. The proposed method was developed with the objective of improving perceptual reconstruction quality while maintaining temporal consistency across video frames. To achieve this, the framework integrates RAFT-based temporal alignment, multi-frame fusion and adversarial learning within a unified architecture. The experimental results on REDS_val and VID4 show that the proposed approach reconstructs HR frames with better texture sharpness and structural preservation than the reconstruction-only baseline. The integration of adversarial and perceptual learning reduces the over-smoothing commonly seen in conventional reconstruction-based methods, and flow-guided alignment helps maintain continuity between neighboring frames.

These results were obtained with a fixed, comparatively small network and a limited training budget, well below the schedules used for large-scale frameworks such as VRT and BasicVSR++. Within this setting, the ablation study isolates the contribution of each component: adversarial perceptual learning gives the largest reduction in LPIPS (0.192 to 0.178) for a small additional cost in PSNR within the ablation configurations. The proposed framework does not reach the PSNR of larger, more heavily trained architectures, but it offers a lightweight and stable approach to perceptually enhanced video super-resolution, and a basis for future work involving advanced temporal transformers, deformable attention mechanisms, and real-time VSR architectures under larger training budgets.

Author Contributions

Conceptualization, Methodology, Experimentation, Data Curation, Results Analysis, Table and Figure Preparation, Writing: Ketaki Sawashe

Supervision, Validation, Methodology Review, Visualization, Resources, Writing, Review & Editing: Bageshree Pathak & Supriya Mangale.

All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Acknowledgments

The authors would like to thank MKSSS's Cummins College of Engineering for Women, Pune, for providing the facilities and support required to carry out this research.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Generative AI Statement

During the preparation of this work, the authors used OpenAI ChatGPT to assist in improving the presentation, organization, and academic writing of the manuscript. The research concept, system design, methodology, implementation, experimental results, and conclusions are the authors' original work. After using this tool, the authors reviewed and edited all AI-assisted content as needed and take full responsibility for the content of this publication.

References

1. A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video Super-Resolution with Convolutional Neural Networks", *IEEE Transactions on Computational Imaging*, Vol. 2, No. 2, pp. 109–122, 2016.
2. J. Caballero, C. Ledig, A. P. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, "Real-Time Video Super-Resolution with Spatio-Temporal Networks and Motion Compensation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2848–2857, 2017.
3. X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-Revealing Deep Video Super-Resolution", *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4472–4480, 2017.
4. M. S. M. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-Recurrent Video Super-Resolution", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6626–6634, 2018.
5. Y. Jo, S. Wug Oh, J. Kang, and S. J. Kim, "Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3224–3232, 2018.
6. J. Liang, H. Fan, S. Xiang, Y. Liu, and R. Timofte, "VRT: A Video Restoration Transformer," *IEEE Transactions on Image Processing*, Vol. 33, pp. 2171–2182, 2024.
7. C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4681–4690, 2017.
8. X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," In: *Proc. of European Conference on Computer Vision (ECCV) Workshops*, pp. 63–79, 2018.
9. Y. Huang, W. Wang, and L. Wang, "Bidirectional Recurrent Convolutional Networks for Multi-Frame Super-Resolution," In: *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 28, pp. 235–243, 2015.
10. Y. Tai, J. Yang, and X. Liu, "Image Super-Resolution via Deep Recursive Residual Network," In: *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3147–3155, 2017.
11. M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent Back-Projection Network for Video Super-Resolution," In: *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3897–3906, 2019.
12. K. C. K. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond," In: *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4947–4956, 2021.
13. K. C. K. Chan, S. Zhou, X. Xu, and C. C. Loy, "BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment," In: *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5972–5981, 2022.
14. X. Wang, K. C. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video Restoration with Enhanced Deformable Convolutional Networks," In: *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.
15. Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "TDAN: Temporally-Deformable Alignment Network for Video Super-Resolution," In: *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3360–3369, 2020.
16. H. Wang, D. Su, C. Liu, L. Jin, X. Sun, and X. Peng, "Deformable Non-Local Network for Video Super-Resolution," *IEEE Access*, Vol. 7, pp. 177734–177744, 2019.
17. X. Ying, L. Wang, Y. Wang, W. Sheng, W. An, and Y. Guo, "Deformable 3D Convolution for Video Super-Resolution," *IEEE Signal Processing Letters*, Vol. 27, pp. 1500–1504, 2020.
18. J. Lin, Y. Huang, and L. Wang, "FDAN: Flow-Guided Deformable Alignment Network for Video Super-Resolution," *arXiv preprint arXiv:2105.05640*, 2021.

19. T. H. Kim, M. S. Sajjadi, M. Hirsch, and B. Schölkopf, "Spatio-Temporal Transformer Network for Video Restoration," In: Proc. of European Conference on Computer Vision (ECCV), pp. 106–122, 2018.
20. J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image Restoration Using Swin Transformer," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1833–1844, 2021.
21. J. Cao, Y. Li, K. Zhang, and L. Van Gool, "Video Super-Resolution Transformer," arXiv preprint arXiv:2106.06847, 2021.
22. C. Liu, H. Yang, J. Fu, and X. Qian, "Learning Trajectory-Aware Transformer for Video Super-Resolution," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5687–5696, 2022.
23. Z. Geng, L. Liang, T. Ding, and I. Zharkov, "RSTT: Real-Time Spatial Temporal Transformer for Space-Time Video Super-Resolution," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17441–17451, 2022.
24. D. Su, H. Wang, L. Jin, X. Sun, and X. Peng, "Local-Global Fusion Network for Video Super-Resolution," IEEE Access, Vol. 8, pp. 172443–172456, 2020.
25. J. Tang, C. Lu, Z. Liu, J. Li, H. Dai, and Y. Ding, "CTVSR: Collaborative Spatial–Temporal Transformer for Video Super-Resolution," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 34, No. 6, pp. 5018–5032, 2024.
26. C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4681–4690, 2017.
27. X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 63–79, 2018.
28. X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1905–1914, 2021.
29. M. Chu, Y. Xie, J. Mayer, L. Leal-Taixé, and N. Thuerey, "Learning Temporal Coherence via Self-Supervision for GAN-Based Video Generation," ACM Transactions on Graphics, Vol. 39, No. 4, pp. 75:1–75:12, 2020.
30. A. Chadha, J. Britto, and M. M. Roja, "iSeeBetter: Spatio-Temporal Video Super-Resolution Using Recurrent Generative Back-Projection Networks," Computational Visual Media, Vol. 6, No. 3, pp. 307–317, 2020.
31. P. Yi, Z. Wang, K. Jiang, J. Jiang, T. Lu, and J. Ma, "A Progressive Fusion Generative Adversarial Network for Realistic and Consistent Video Super-Resolution," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 44, No. 5, pp. 2264–2280, 2022.
32. Fahmy, M. Sulaiman, Z. Shehabeldin, M. Barakat, M. El-Naggar, D. Hussein, M. Youssef, and H. M. Eraqi, "RBPGAN: Recurrent Back-Projection GAN for Video Super Resolution," Journal of Artificial Intelligence and Autonomous Intelligence, Vol. 1, No. 1, pp. 32–49, 2024.
33. A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "FlowNet: Learning Optical Flow with Convolutional Networks," Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2758–2766, 2015.
34. E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2462–2470, 2017.
35. Z. Teed and J. Deng, "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow," Proceedings of the European Conference on Computer Vision (ECCV), Vol. 12347, pp. 402–419, 2020.
36. W. Bao, W.-S. Lai, X. Zhang, Z. Gao, and M.-H. Yang, "MEMC-Net: Motion Estimation and Motion Compensation Driven Neural Network for Video Interpolation and Enhancement," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, No. 3, pp. 933–948, 2021.